4

RADC-TR-89-136
Final Technical Report
September 1989

OPTICAL SHARED MEMORY

University of California, Davis

Stephen T. Kowel, Norman Matloff, C. Eldering, T. Schubert, M. Loving

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441-5700

This report has been reviewed by the RADC Public Affairs Division (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-89-136 has been reviewed and is approved for publication.

APPROVED: Robert L. Kaminski

ROBERT L. KAMINSKI
Project Engineer

APPROVED: Raymond P. Urtz Jr

RAYMOND P. URTZ, Jr.
Technical Director
Directorate of Command & Control

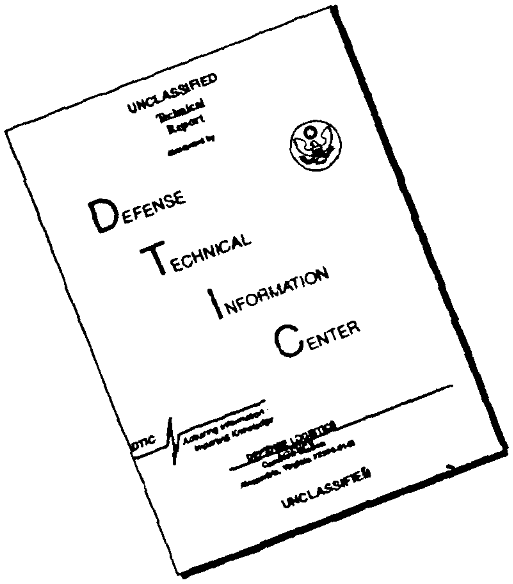
FOR THE COMMANDER:

BILLY G. OAKS
Directorate of Plans & Programs

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (COTC) Griffiss AFB NY 13441-5700. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

DISCLAIMER NOTICE



UNCLASSIFIED

Technical

Report

distributed by

DEFENSE

TECHNICAL

INFORMATION

CENTER

DTIC Acquiring Information Imparting Knowledge

DEFENSE LOGISTICS AGENCY

Cameron Station

Alexandria, Virginia 22304-6145

UNCLASSIFIED

THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS
N/A

2a. SECURITY CLASSIFICATION AUTHORITY
N/A

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE
N/A

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
N/A

5. MONITORING ORGANIZATION REPORT NUMBER(S)
RADC-TR-89-136

6a. NAME OF PERFORMING ORGANIZATION
University of California, Davis

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Rome Air Development Center (COTC)

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Electrical Engineering and
Computer Science
Davis CA 95616

7b. ADDRESS *(City, State, and ZIP Code)*
Griffiss AFB NY 13441-5700

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Rome Air Development Center

8b. OFFICE SYMBOL *(If applicable)*
COTC

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F30602-81-C-0193

8c. ADDRESS *(City, State, and ZIP Code)*
Griffiss AFB NY 13441-5700

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
63728F	2529	01	P9

11. TITLE *(Include Security Classification)*
OPTICAL SHARED MEMORY

12. PERSONAL AUTHOR(S)
Stephen T. Kowel, Norman Matloff, C. Eldering, T. Schubert, M. Loving

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM Mar 88 TO Dec 88

14. DATE OF REPORT *(Year, Month, Day)*
September 1989

15. PAGE COUNT
48

16. SUPPLEMENTARY NOTATION
N/A

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
12	07	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Optical Shared Memory
Electro-Optical Multiprocessors
Optical Computing

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The objective of this effort was to investigate and evaluate OPTIMUL, an Optical Interconnect for Multiprocessor Systems, which could be used to interconnect a range of granularity of processors and memories for access to very large data/knowledge bases (10**12-10**16 bytes). The final report details results of the architecture simulation, and discusses the operation of the system. The report also includes designs for an Optically Writeable Ram Cell (OWRC), a Differential Amplifier/Static Ram Cell (SRAM), and a Balanced Receiver.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Robert L. Kaminski

22b. TELEPHONE *(Include Area Code)*
(315) 330-2925

22c. OFFICE SYMBOL
RADC (COTC)

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

1. Introduction

Work has been concentrated in three areas; system design and applications, memory design, and transmitter design. Goals for each of these areas have been determined and work has progressed to provide detailed simulations of the ***OFTIMUL*** (for Optically Interconnected Multiprocessor) system and demonstrate its usefulness in a number of applications. A preliminary design for the system was completed and plans for fabrication of a prototype system were developed.

Professor Kowel attended the 1988 ACM International Conference on Supercomputing in July, and presented the paper, entitled " ***OPTIMUL***: An Optical Interconnect for Multiprocessor Systems", included in the Appendix. One of the major invited talks, by Carl Ledbetter, President of ETA, dealt with the challenges of obtaining a factor of 10 improvement in supercomputer performance. He showed that fundamental physical constraints as well as practical fabrication problems rule out success by traditional technological paths. He mentioned two possible paths to gain significant improvements - software, and hybrid electronic/optical systems. The work done during this reporting period encourages us in our belief that we have a promising solution, based on both categories.

2. System Design and Applications

Both an optical read and an optical read/write system have been evaluated for their potential increase in speed in a multiprocessor environment. It may turn out that the potential increase in speed is greatest in loosely coupled systems. Database applications were studied extensively as an application which can benefit greatly from the use of optical memory interconnect as proposed in ***OPTIMUL***. The use of the ***OPTIMUL*** system in projection, sort and join operations were studied. Select and project operations lend themselves easily to simple multiprocessor operations and optical interconnect will allow for reception of the partitioned task without contention and subsequent delay. Sorting and joining operations have also been studied and algorithms utilizing optical interconnects were developed. The results of this portion of the work have been accepted for the Eighth Annual IEEE International Phoenix Conference on Computers and Communications to be held in Scottsdale, AZ, in March, 1989. A copy of this manuscript is included in the Appendix.

In addition to the work performed in the area of relational database applications for the ***OPTIMUL*** system, the following tasks were undertaken:

1. Identification of a particular computational problem which will benefit most greatly from the ***OPTIMUL*** technology. Possible problems include pattern recognition and classification, weather prediction, or expert system tasks.

2. Algorithm development for solution of the identified problem using an optically interconnect multiprocessor system.

3. Simulation of the system using various sized memories and transfer rates, and various system configurations.

2. Memory Design

A preliminary design of an Optically Writeable Ram Cell (OWRC) has been completed and simulations of the device performed. The device is essentially a fast static ram cell with the capability of being written to optically. Electronic writes may also be performed in which case the device acts as a standard memory device. The circuit uses reverse-biased photodiodes to act as optical detectors; these detectors are modeled as current sources which generate current in proportion to the amount of illumination. The most important feature of this design is that the device acts as a differential detector and can determine small differences between a reference beam and the information beam. In this way the memory information may be transmitted to the receiver without full modulation of the incident beam. As discussed in this report, simulations have shown that with a modulation level of less than 1%, the memory information can be received in 10ns.

2.1 Basic Operation

The OWRC consists of three major functional blocks; 1) input circuitry, 2) the differential amplifier/SRAM cell, and 3) the output circuitry. It has two data inputs and requires four controlling clocks. Two of these clocks require the inverse signal to drive the P-channel devices. The complements can be generated by adding inverters to the cell but to maintain a minimum size they have been assumed to be provided. Thus there are a total of 8 inputs (6 for clock signal and 2 data). The results of circuit simulations of the devices are shown in Figures 4 and 5.

2.2 Detailed Operation

OWRC employs differential optical inputs to maximize resolution and to minimize the constraints on the optical system which will be supplying the input signals. The input circuit is shown in Figure 1a and the corresponding circuit model shown in Figure 1b. The positive and negative inputs are identical in every way. The input circuitry drives the differential amplfier stage which can be viewed as a large capacitance.

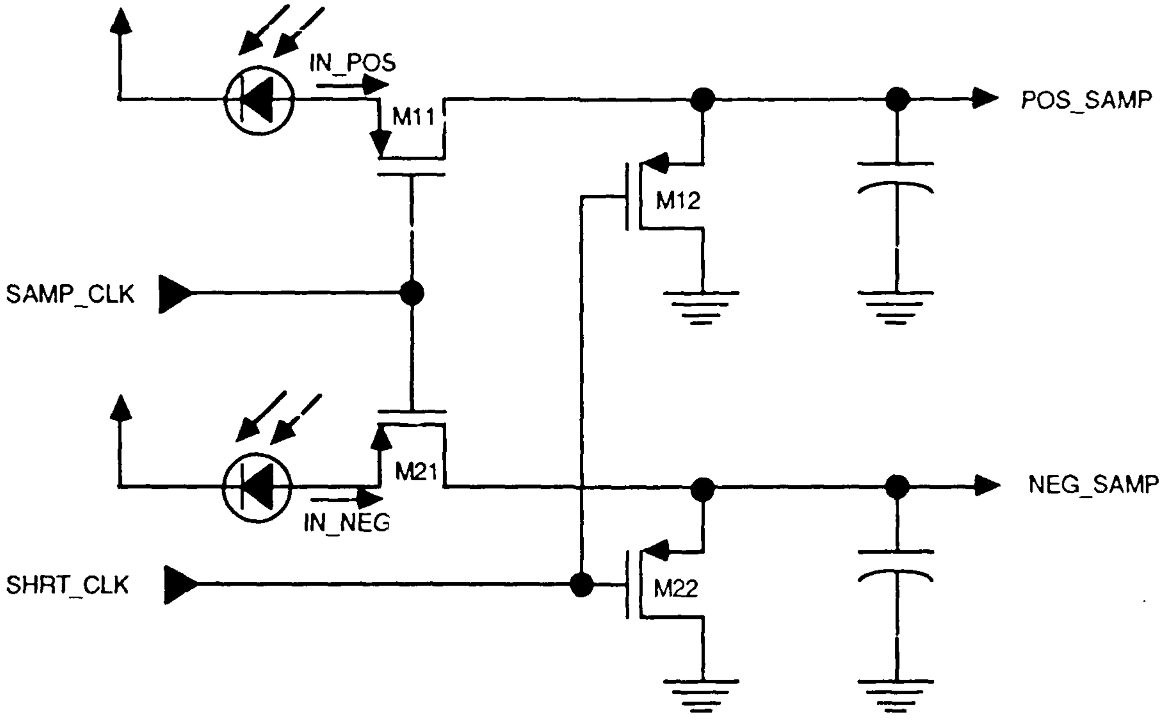
The optical receivers are reverse-biased diodes which will conduct a current proportional to its illumination. Since the diodes may be under contstant illumination., the nodes IN_POS and IN_NEG will normally be at 5 volts. To operate the circuit, the nodes POS_SAMP and NEG_SAMP must first be discharged to ground so that they will start charging from the same level. This is done by asserting SHRT_CLK which drives the gates of M11 and M21 to charge POS_SAMP and NEG_SAMP. The final voltages are determined by the illumination and the duration of SAMP_CLK. With a different amount of illumination on the two diodes (representitive of a one in memory), the POS_SAMP and NEG_SAMP nodes will charge at different rates and will thus be at different final voltages when SAMP_CLK is negated.

2.3 Differential Amplifier / Static Ram (SRAM) Cell

The differential amplifier, as illustrated in Figure 2, consists of two cross coupled CMOS inverters with two additional FETs to allow the application and removal of power to the two inverters. Once the sample has been taken and SAMP_CLK has been negated the difference may be evaluated by asserting the evaluation clocks EVAL_CLK and NOT_EVAL. As can be seen in Figure 2, EVAL_CLK drives an N-channel FET which connects the inverters to ground and NOT_EVAL drives the complementary P channel FETS which connects the inverters to power. When power and ground are applied to the inverters they amplify the difference between the two sample nodes (inputs to the inverters) and settle with the higher one at 5 volts and the lower one at ground. It is important that the input circuitry supply current such that these nodes charge to a level V_0 which is constrained such that $(V_{dd} - V_{thp}) > V_0 > (V_{ss} + V_{thp})$. Failure to meet this condition will result in either the P or N channel devices in both inverters being off when evaluation starts causing unpredictable results. EVAL_CLK is held high as long as it is desired to maintain the data in the RAM cell.

2.4 Output Stage

It is important that the capacitance from the nodes POS_SAMP and NEG_SAMP be closely matched since deviation from perfect matching will result in degradation of the resolution of the difference amplification. With this in mind, the output of the OWRC is also differential. This is for capacitance matching purposes. The output stage consists of a simple CMOS pass gate which is shown in the diagram of the complete circuit (Figure 3). Data may be read out of the RAM cell anytime after it has settled by asserting ENL_CLK and INV_ENL, thus connecting OUTPUT (INV_OUT) to POS_SAMP (NEG_SAMP).



a

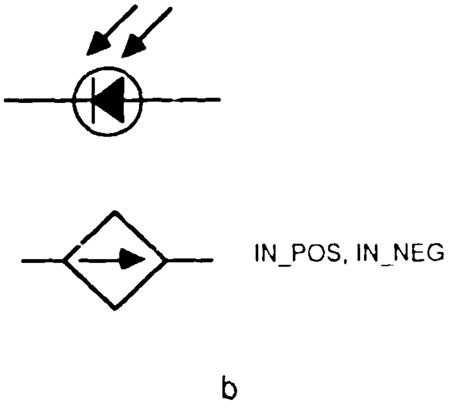


Figure 1. Input circuit (a) and model for photodiodes (b).

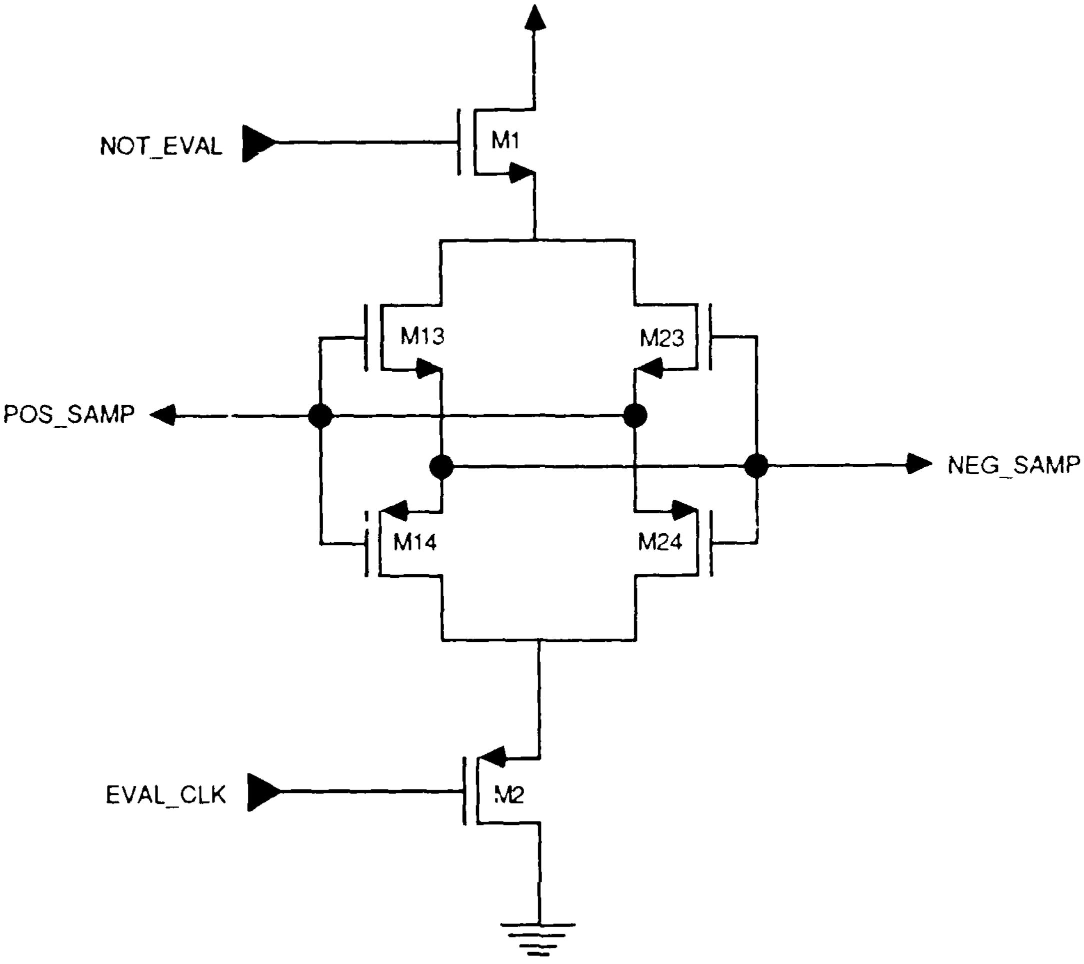


Figure 2. Differential Amplifier /SRAM Cel!

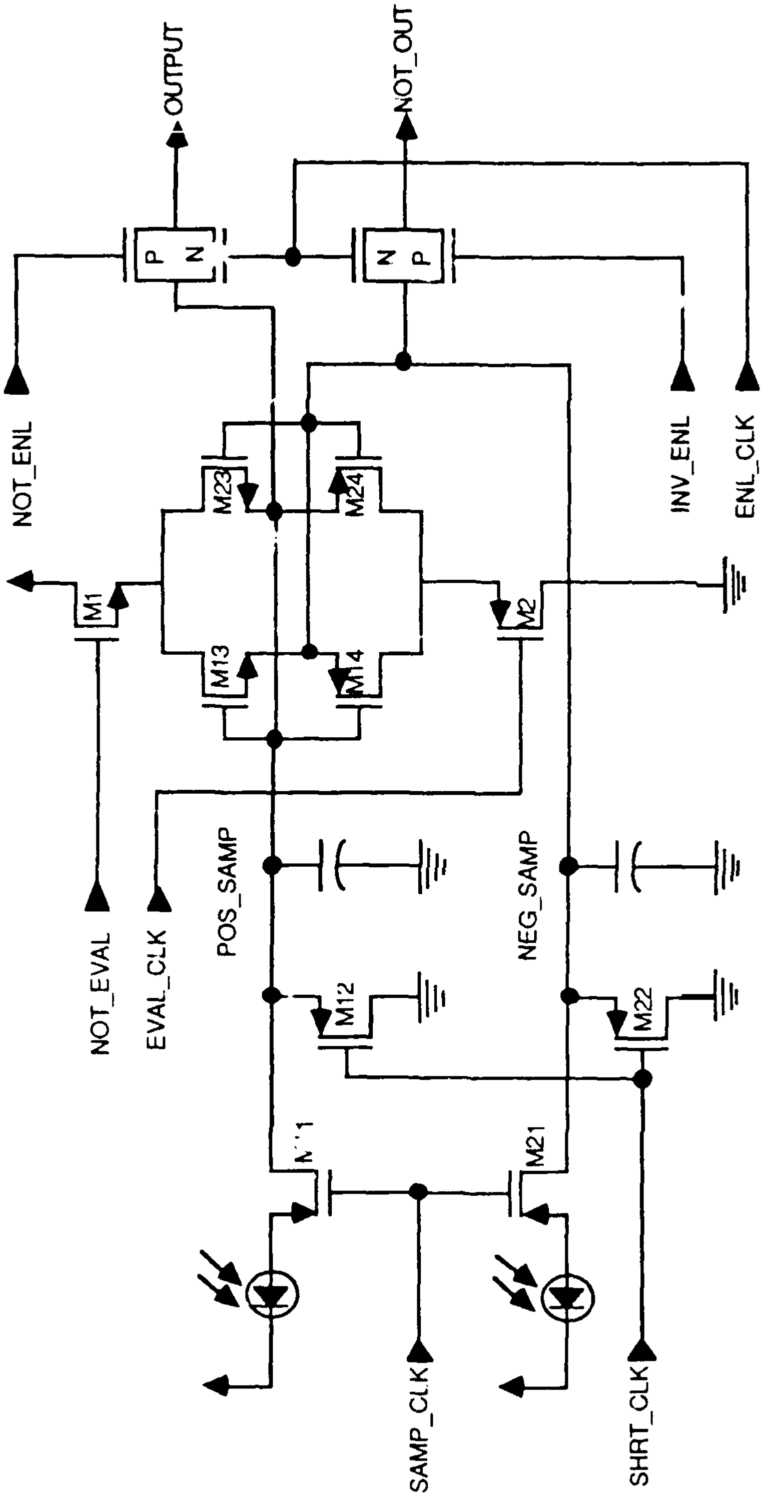


Figure 3. Schematic for entire Optically Writeable Ram Cell (OWRC)

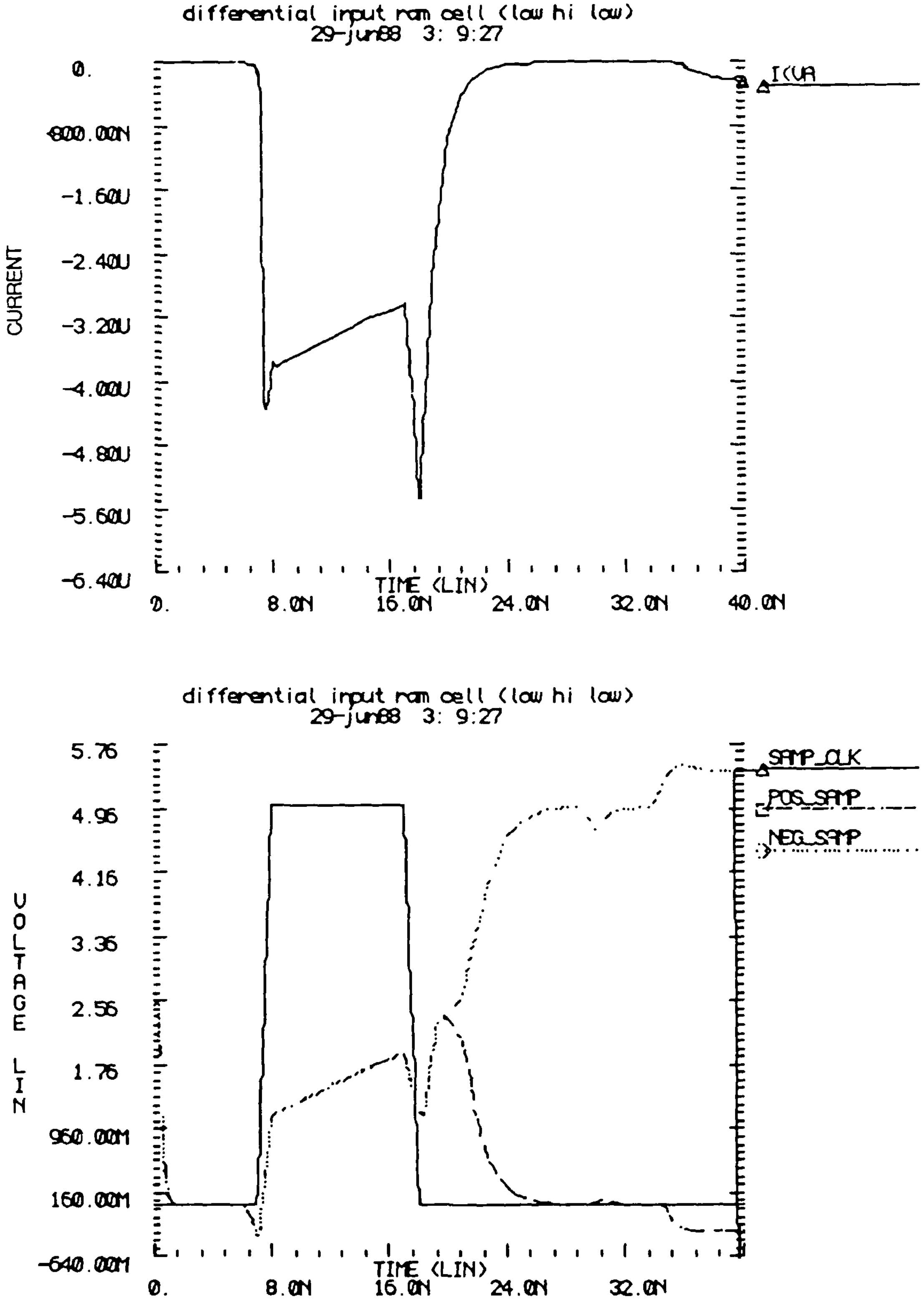


Figure 4. Circuit simulations performed using Spice which demonstrate the response of the OWRC (lower simulation) to an optically induced current of approximately 4μ amperes (top simulation).

2.5 Balanced Receiver Design

During the last period of this contract, the balanced receiver was analyzed as a possible input structure for the memory. The two photodiodes act as a differencing element for the optical signals received as shown in Figure 5. By using this as an input to the memory circuit described previously, it should be possible to reduce the complexity and thus the real estate requirements for the receiver. The use of a coherent receiving system was also investigated, as shown in Figure 6. The advantage of this receiving system is that it does not require polarizing filters and has better theoretical sensitivity, but requires more sophisticated optics to create the interference.

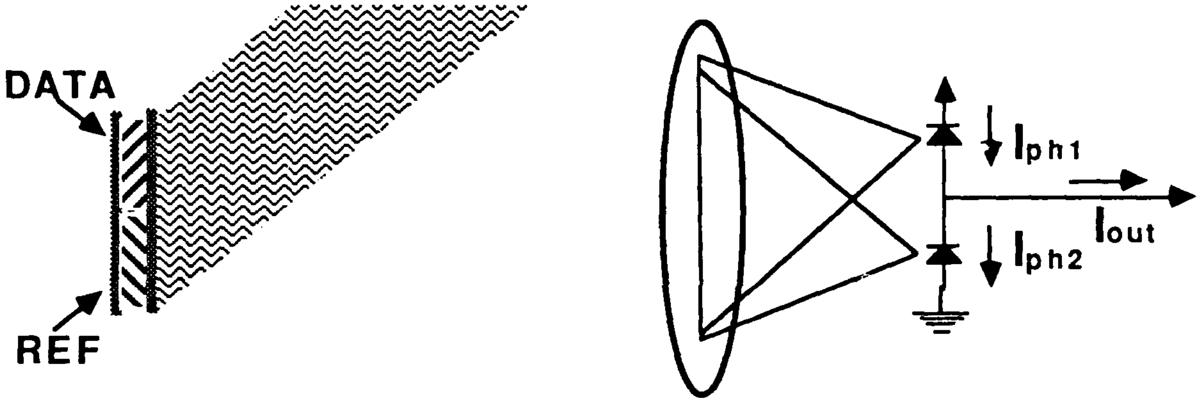


Figure 5. Balanced receiver for detection of amplitude modulated signals. Reference cell transmits the dark level while the data pixel transmits the memory information.

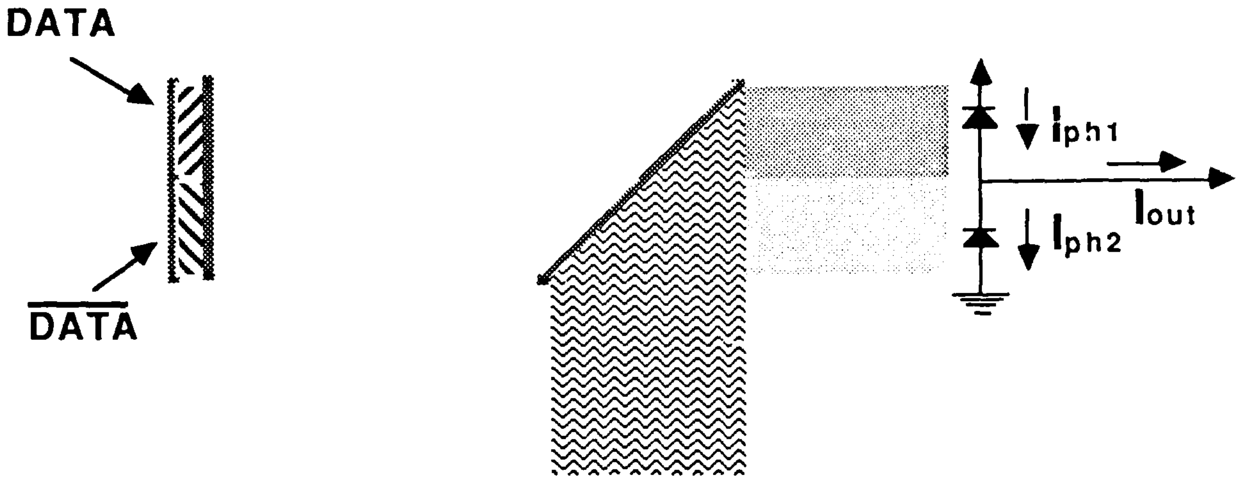


Figure 6. Balanced receiver used as a coherent detector. In this case no polarizing filters are necessary and the differential phase shift between the two pixels is measured. This configuration has the highest theoretical sensitivity but requires vibration-free optics and a coherent source.

3. Optical Design

Basea on the preliminary design of the optical receiving elements, an optical budget can be estimated for the system. Figure 7 illustrates the incident power on a transmitting element and the subsequent propagation and losses in a system containing 8 receiving arrays.

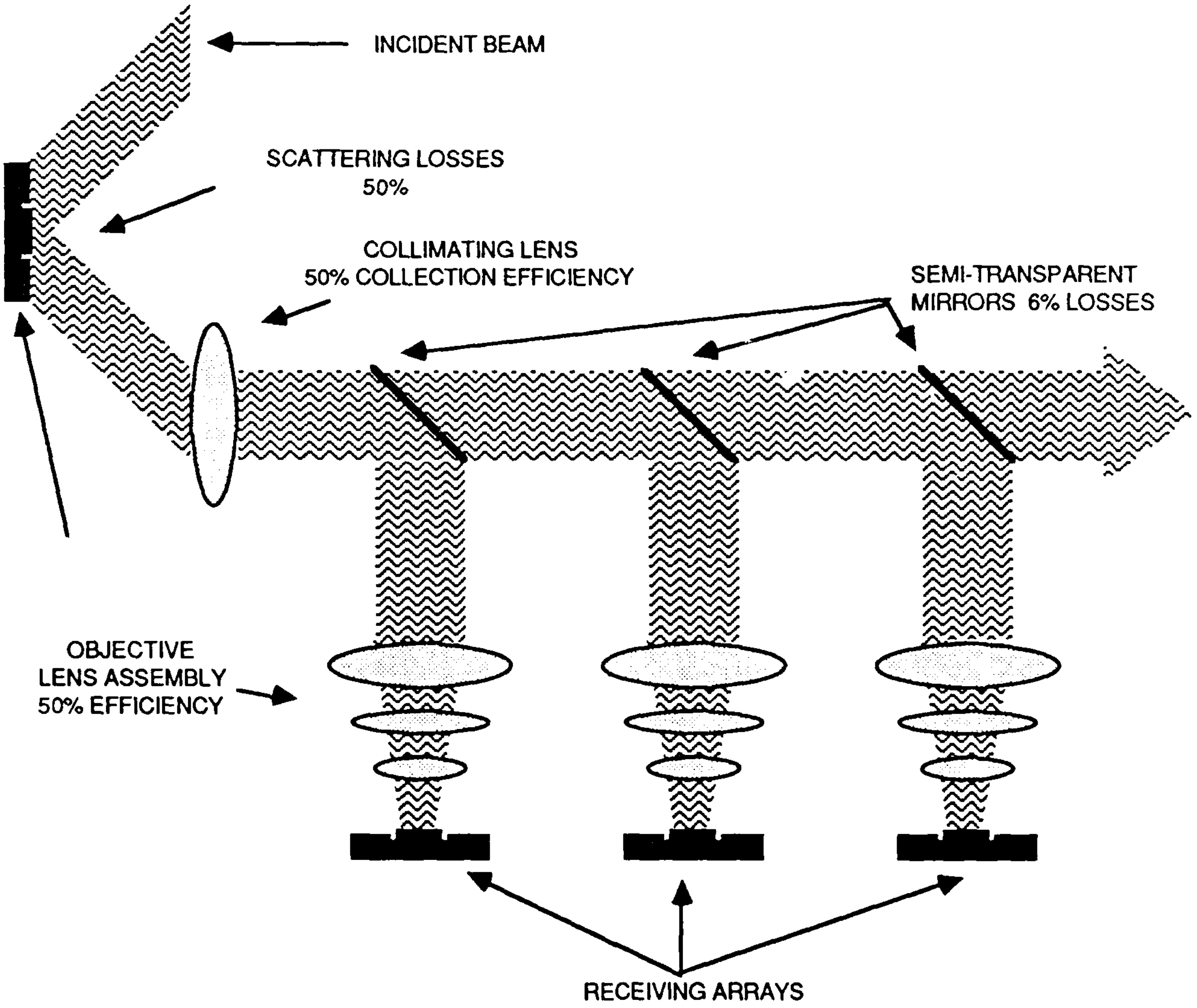


Figure 7. Optical Schematic for Budget Calculations.

Detailed calculations based on the specifications of available CCD devices as receivers and ferroelectric liquid crystals as the modulation coating have been made. They reveal that 1 Watt of input power is sufficient to drive 64 processors from one 64Kb shared memory, assuming 'no repeator' architecture. This optical budget can certainly be provided by a modest gas laser, or by an incandescent source. For a thin solid film coating, the estimation is more difficult. Our curent AZO-DYE etalons provide only 0.01% modulation, compared to nearly 100% for the liquid crystal films. Of course, we expect to make far better etalons with better dyes as the work continues. With an improvement of 100, we should be able to design electronics capable of discriminating the two switched levels.

4. Optical Budget

Based on the design of a multiple image system based on beam-splitters (as shown in the previous quarterly report) an analysis of the total modulated optical power needed as a function of the switching current per bit was calculated. The total optical input power required is found to be

$$P_0 = \frac{n}{s \cdot c \cdot o \cdot m^n \cdot D} \cdot i_s$$

where

P_0 = required optical input power

n = number of receiving arrays

s = scattering loss factor (assumed to be 0.5)

c = collimating lens loss factor (assumed to be 0.5)

m = mirror loss factor (assumed to be 0.92)

D = receiver detectivity (assumed to be 0.3A/W)

i_s = switching current

Figure 8 reveals several cases. For the simulations shown below the modulation was considered to be 100% efficient. For modulation efficiencies less than 100% the required optical power will increase linearly with the decrease in modulating efficiency. As illustrated in the graph, if switching currents on the order of 10 nA are sufficient for switching the memory bits (with a bit error rate of < 10^{-11}) the system will require less than 10 Watts of optical input power, even with 32 processors. While it is possible to obtain lasers with this much continuous power, it will be more feasible to use filtered incandescent light as a source. Filtering the light from a broadband source will provide an inexpensive yet strong (>10W) source of light. It is interesting to note that the Fabry-Perot etalons only have an effective path length on the order of 500µm and thus the coherence length of the light needs to be on the order of 1mm. Since ordinary discharge lamps have coherence lengths on the order of several mm, is should be possible to obtain a powerful yet inexpensive light source for thin film modulators.

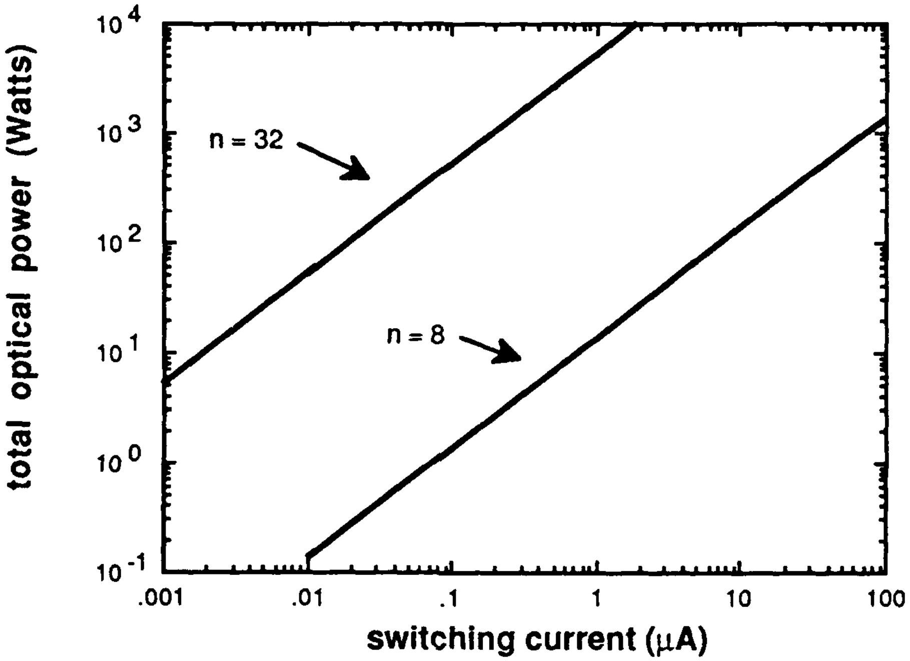


Figure 8. Switching current versus optical power.

5. Summary of a Case Study:

Sorting Application

Assumptions

- loosely coupled system (ring topology)
- *OPTIMUL* transmission speed 500 ns
- non- *OPTIMUL* speed 50MBits/sec
- memory transfer size from 16K 32 bit integers to 256 K 32 bit integers
- effective *OPTIMUL* transmission 1 GBits/sec to 16 GBits/sec
- each processor : VAX 8600 equivalent

Results: 128K integer array

# of processors multiprocessor system	speedup compared to conventional
2	1:1.05
8	1:1.17
32	1:1.78
128	1:2.68

Multiprocessor Interconnect Simulation

Assumptions

16 processors

8 memories: each 512Words, 20 Bits/Word

Total Memory Space = 4K

O(n) problem with 84% reads and 16% writes

Adds 256 numbers: each processor adds 16 numbers

MIC -1 processors

Assume memory cycle= microcycle

Results:

connection network	ratio to ideal	read ratio	write ratio
ideal	1.0	1.0	1.0
single bus	2.15		
crossbar	1.37		
OPTIMUL-1	1.07	1.1	1.4

Conclusions

- Speedup achieved is dependendent on relationship between computation complexity and communication complexity

- Some applications may show a decrease of at least an order of magnitude in run time using *OPTIMUL* technology

- Applications previously not feasible for multiprocessor systems could be run on an *OPTIMUL* system

- Ferroelectric liquid crystals will be suitable for a protype system

- Modulation has been demonstrated in solid nonlinear films and IC surface modulators appear feasible.

6. Conclusions

During the contract work has progressed in all areas of the program. By performing simulations at both the system and circuit level, we are able to predict the overall performance of an ***OPTIMUL*** system and allow for the development of a preliminary design. This design allows for implementation in the immediate future using available materials such as fast liquid crystals but will be applicable to other technologies being developed such as polymeric electro-optical thin film materials.

Concerning the tasks which lie ahead of us, the following should be mentioned as the most crucial ones:

a. Further investigation of hardware/software methods for interprocess and interprocessor communication.

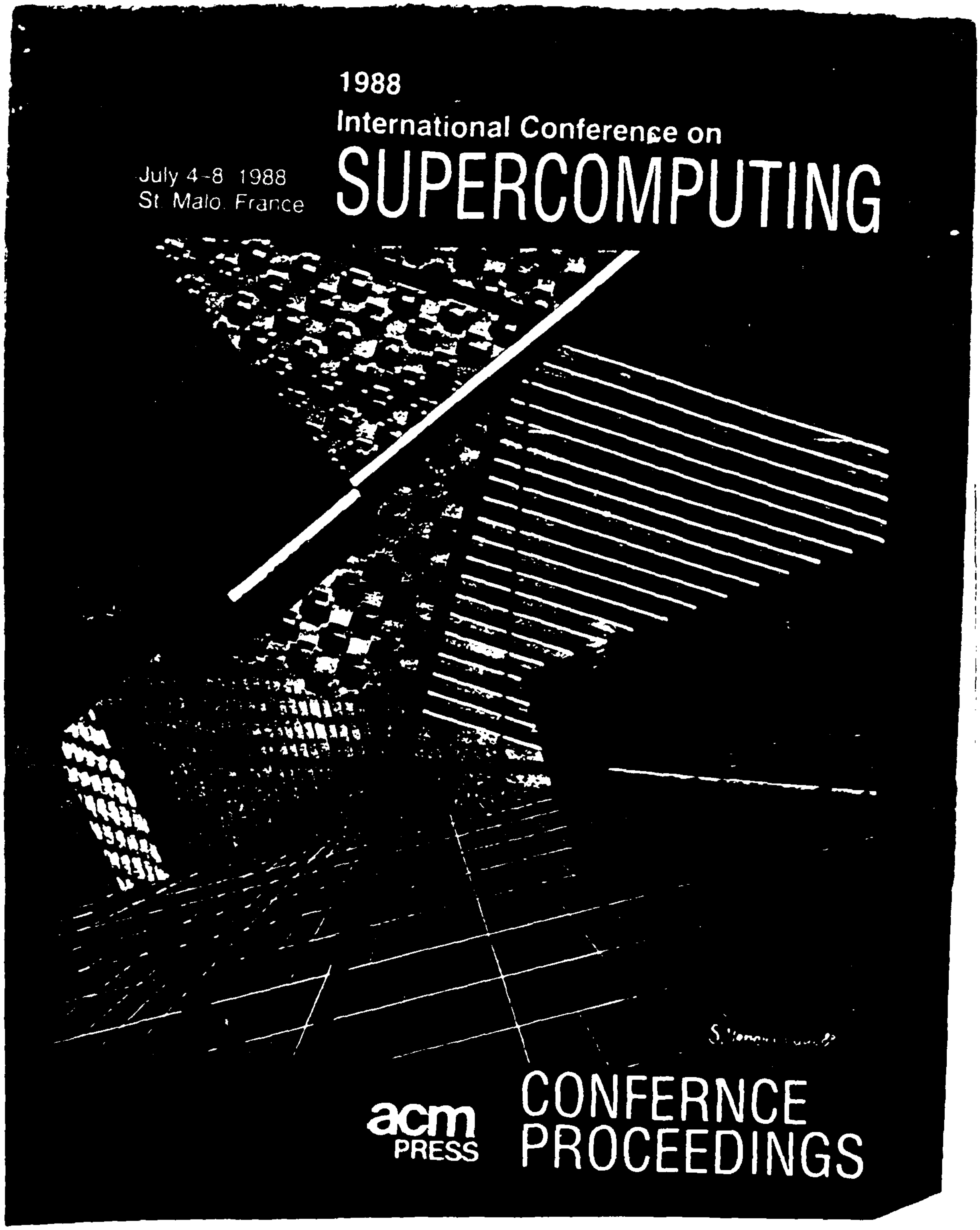
b. Investigation into topologies (tree, ring, grid, etc.) in order to determine how best to exploit the tremendous communication bandwidth of ***OPTIMUL***.

This program has received considerable visibility during the contract period. In addition to the two reviewed conference papers (1988 International Conference on Supercomputing; 1989 IEEE International Phoenix Conference on Computers and Communications), further evidence of the value of the effort can be found in the fact that the US Office of Trademarks and Patents has notified us of the allowance of a patent application (Electro-Optical Interface) submitted previously to the award of this contract.

The calculations included in this report demonstrate that ***OPTIMUL*** can be a viable computer technology and that research and development should be vigorously pursued.

Appendix

Attached are copies of the two papers completed under this contract.



1988

International Conference on

July 4-8 1988
St. Malo, France

SUPERCOMPUTING

acm
PRESS

CONFERNCE
PROCEEDINGS

OPTIMUL: An Optical Interconnect for Multiprocessor Systems

Norman Matloff, Stephen Kowel and Charles Eldering
Department of Electrical Engineering & Computer Science
University of California at Davis

Corresponding Author: Norman Matloff matloff@iris.ucdavis.edu

Abstract. An optical interconnect is proposed for multiprocessor systems, of both the tightly and loosely coupled types. This interconnect solves the problem of contention for memory and interconnect in the tightly coupled case, and the problem of network bottleneck in the loosely coupled case.

1. Introduction

Multiprocessor (MP) systems, consisting of p interconnected but independent processors, have the *potential* for a speedup factor of p in computational power. However, a long-standing problem has been that this potential has not been realizable, due to the overhead of processor-memory and/or processor-processor communication. This has been the case for both types of MP systems which are usually considered.

Tightly-Coupled (TC) Systems

The very significant overhead in TC systems takes the form of contention for the shared central memory M_{cen}, and for the processor-memory interconnect. The latter problem is exacerbated by the fact that the expense of full crossbar switches results in the use of other networks for which there is even more processor contention for the interconnect, e.g. *Ω-nets* (Hwang and Briggs, 1984). For small values of p, use of caches can be effective, but the efficiency decreases with p (Wilson, 1987). Furthermore, it has recently been discovered that access to interprocess synchronization variables in shared memory worsens this problem tremendously (Pfister and Norton, 1986).

TC systems have been used mainly for applications in which the parallelism is fine-grained, requiring frequent interprocessor communication. The effects of the communications overheads described above, though, have limited TC systems to rather small numbers of processors.

Loosely-Coupled (LC) Systems.

In LC systems, there is no shared memory, but there is still communications overhead of another kind. The processors communicate with each other through a network. Bandwidth limitations on this interconnection network present very substantial overhead. For example, intercluster accesses in the Cm* machine were a factor of 9.7 times slower than accesses to local memory (Hwang and Briggs, 1984). The effect of this very high overhead has been that LC systems have been used mainly for coarse-grained applications, in which the low frequency of interprocessor communication implies that this overhead may not be a significant factor.

In both the TC and LC settings, another significant problem is the severe restrictions resulting from pin limitations. Even channels of very high bandwidth, such as those constructed from optical fibers, would not solve the problem arising from the fact that there are only a few data pins but thousands or even millions of bits in a memory chip.

The entire history of the development of MP technology has been dominated by the search for solutions to these problems (Siewiorek *et al*, 1982; Hwang and Briggs, 1984; Agrawal, 1986). Essentially, no completely satisfactory solutions have been found. For example, after Cray Research, Inc. released the Cray X-MP, an MP version of the Cray-1 supercomputer recently, a number of investigations (Bailey, 1987; Cheung and Smith, 1986; Chen and Lange, 1986) quickly showed the system to suffer from slowdowns due to both contention for shared memory and contention for the network which connects the processors to that memory, just as with all the earlier MP systems.

© 1988 ACM 0-89791-272-1/88/0007/0016 $1.50

Perhaps an even more dramatic example is the S-1 a TC MP system developed at Lawrence Livermore National Laboratories Hwang and Briggs. 1984. Throughout the period of development of this system, it was hailed as one of the most advanced MP projects in existence. However, recently the project was discontinued, in spite of all the favorable publicity, and the very extensive funds expended Bruner, 1987. One of the primary reasons given for the discontinuation was that the project engineers had found that the contention for shared memory in the system would be much greater than they had anticipated. They are now beginning work on a completely new design.

We will present here a radically new interconnect method which will solve these problems, and have other advantages as well:

(a) The new interconnect will be usable for both fine-grained and coarse-grained types of applications. Moreover, it could be applied to build systems which are equally effective on both of these application types, with no reconfiguration time. Such systems would then also work well for "medium-grained" applications, thus recognizing that the fine-grained and coarse-grained concepts are merely two extremal representatives in a broad range of problems having varying degrees of frequency of interprocessor communication.

(b) It will solve the long-standing problems of contention *for memory* and for the processor memory switch in TC systems. There will be absolutely no queueing delay for read access to shared memory.

(c) In LC *systems*, it will enable a truly dramatic improvement in interprocessor communications bandwidth, and again totally eliminate contention for the interconnect switch.

(d) Although *there will still be physical* limitations on the size of p, such limits should be far less constraining than those of existing systems with conventional nonoptical processor *memory interconnects*.

(e) Our approach should also be superior to other optical processor/memory interconnects which have been proposed, e.g. optical crossbars Bell, 1986, Hutcheson *et al.* 1987. For example, an optical crossbar switch of size as large as 256 x 256 has been anticipated, but even this would only allow 256 simultaneous bits to be transmitted; by contrast, under our approach, the entire contents of a chip can be transmitted simultaneously, i.e. thousands or even millions of bits can be sent in parallel. Note that this also implies that the pin-limitation problem is also eliminated, which is a problem even in those architectures which have been proposed based on an optical fiber interconnect.

Our name for this new interconnect is *OPTIMUL*, an acronym for Optical Multiprocessor Interconnect. The central feature is an optical processor-memory channel, which will allow simultaneous access of a memory chip, where the word "simultaneous" is meant *both* with respect to all bits in the chip, *and* with respect to all processors. In other words, all processors can simultaneously read the entire contents of a chip with no interference at all. Write access is of course restricted to a single processor at a time, but it still is simultaneous across bits in the chip, i.e. an entire chip can be written in one access. This optical channel is described in Section 2, and then MP system architectures utilizing it will be proposed in Section 3. Section 4 will then present some implementation details.

2. A New Optical Memory Access Channel

Consider devices $D_1, \ldots, D_k$ which wish to read a memory chip C, in which are stored bits $B_1, \ldots, B_l$. We will report here a technique in which the devices can read from C optically, bypassing the need for using the chip's pins, and which will allow this access to be simultaneous, with respect to both devices D_i and bits B_j (Figure 1).

To achieve this, C will be coated with a thin polymeric film, using a Langmuir/Blodgett (L/B) or other technique Kowel *et al.* 1985] [Kowel *et al.* 1987]. When C is illuminated, e.g. by a laser, the film will cause the reflected beam to be intensity-modulated by the electric fields at each position beneath the film in C. Thus the reflected beam will contain a complete bit map of the contents of C. The beam will be processed by optical apparatus for focusing onto the receivers D_i.

Demodulation of the beam back to storage as electric fields at the receivers is accomplished by the use of photosensitive technology. For example, one possibility is to use ordinary DRAM memories, which have a natural sensitivity to light. This means also that parts of C must be masked from the light, so that illumination of C does not change the contents of bits in C, e.g. only the output portion of a gate can be exposed. CCD or CID arrays are also possibilities for use as demodulators.

In this way, the values stored at all the bits B_j in C can be transmitted optically to the devices D_i, *simultaneously over all subscripts i and j*. Clearly, the simultaneity over i will have highly significant implications for the memory and interconnect contention problems which have plagued TC systems, while the simultaneity over both i and j will have an equally profound impact on the network bandwidth limitation problem in LC systems. Note again that the classical bottleneck arising from limitations on the pins-to-stored-bits ratio is completely bypassed in the approach described here.

Writes to C can be accomplished by reversing the process (or by using separate chips, as in Architecture II in Section 3). E.g. suppose an entity associated with D_i wishes to write to C. At the time the system is built, D_i would be coated with the same type of film as described for C. Then to write to C, the following would be done: The entity places the items to be written into D_i (in most system architectures based on this interconnect technology, the items will already be D_i anyway; see Section 3). Then the illumination to C is turned off, and D_i is illuminated instead. The effect is that the contents of D_i are copied by C.

3. System Architectures

The optical interconnect presented here can be used in a variety of configurations. We discuss two examples in this section.

Architecture I:

This will be the first system to be built. The system will consist of the following (Figure 2):

(a) A system bus S_0

(b) A central shared memory, M_{cent}, consisting of L B-coated chips M_{0j}, j = 1,...,m

(c) p processor memory bus modules. In the i-th of these, a processor P_i will be connected via a local bus S_i to "local memory" $M_{loc\,i}$, consisting of L B-coated memory chips M_{ij}, j = 1,...,m. Note that in spite of the name "local memory," M_{ij} serves primarily not as memory but rather as a receiver for the reflected beam from M_{0j}: P_i reads M_{0j} by reading M_{ij}, which contains an up-to-date copy of M_{0j} at all times, since the illumination of M_{0j} is maintained continuously.

Memory reads are handled optically, in the manner described above. Writes are handled electronically, through the system bus. The system bus also indirectly serves as a mechanism for dealing with the process synchronization problem: Suppose P_a is writing to a shared variable in chip M_{0j} in M_{cent}, and during this time P_b wishes to read that variable. How do we temporarily suppress access by P_b? In an ordinary (nonoptical) system, this is handled through special bus operations which allow a read-modify-write cycle, but that *would not be* available in a totally optical system.

However, since in this architecture we still do have a system bus S_0, the problem can be solved by providing logic which will compare the address portions of S_0 and S_b, and temporarily suppress the signal to the Chip Select pin on M_b, if appropriate. All of this would be transparent to both P_a and P_b, though P_b might "see" a *delayed memory access response*.

Note that we have said nothing here about logical versus physical structure of the shared address space M_{cent}. For example, most existing TC systems use a low-order interleaving scheme for assigning addresses to memory chips, the idea behind such a scheme being to alleviate some of the problems of memory contention. However, there would appear to be no particular advantage to this form of addressing in the optical architectures described here.

Although this system is inhibited somewhat by having optical reads but only conventional writes, the loss may not be very large. Typically reads *dominate* writes by a ratio of 3:1 or more, with some applications (e.g. in-core database searches) having virtually 100% of memory access being in the read mode.

Architecture II:

This is in some sense the other extreme point in a spectrum of possible configurations. The architecture here is designed to most fully exploit the tremendous potential for parallelism provided by the optical technology described in Section 2; the price paid is in extra memory, and possibly extra computing apparatus (i.e. additional lenses and/or mirrors).

The desirable qualities for which we are aiming in this architecture are

- both reading and writing being done optically
- avoidance of using electronic media for access to shared variables and other synchronization problems
- avoidance of delays due to writing, e.g. caused by the need to change the direction of illumination
- avoidance of the need to develop chip technologies which allow both optical read and write access

To accomplish these goals, we have formulated a configuration consisting of the following. There will be p+1 processor memory bus modules (Figure 3), where the i-th module consists of a processor P_i, memory units *(each consisting of many chips, but for simplicity referred* to here as a single "chip") M_{ijk}, and bus S_i. Among the memory units M_{ijk} for Module i, the subscript j ranges from 1 to p for i = 0, while for i > 0, j is equal to i only; the subscript k takes on the values 0 and 1 for any value of i.

Here is how the system works: Modules 1 through p do the computation, while the "leader" Module 0 manages the operation of the system and serves as a data source/collector. More specifically

(a) Any memory chip in the lead module of the labeling form M_{0i0} ($i > 0$) is L B-coated and under constant illumination, and is constantly *read* by the computational processor P_i through P_i's local memory M_{i10}, as in Architecture I. As before, this is accomplished by virtue of the fact that M_{i10} will always have an up-to-date copy of the contents of M_{0i0}.

In this configuration, each computational processor has read access to a different set of chips in the lead memory. (An alternative configuration, which we do not discuss in detail here for the sake of notational simplicity, would be to have *all* read-chips in the lead memory readable by *all* computational processors. This would economize on memory, and would increase speed somewhat in applications in which the same data is to be broadcast to all computational processors. Note, however, that in this configuration, there still would be separate write-chips for each computational processor, i.e. nothing in (b) below would be changed.)

(b) The other lead memory chips, of the labeling form M_{0i1}, are intended to be *written to* by the computational processors P_i ($i > 0$). We will also describe this as P_0 *reading* the memories M_{i11}. This is accomplished by having the corresponding chips in the computational modules, of the labeling form M_{i11}, L B-coated and under constant illumination, so that M_{0i1} will constantly have an up-to-date copy of M_{i11}.

(c) The process synchronization problem will be handled by message passing [Peterson & Silberschatz, 1985]. A computational processor P_i can access a message from the lead processor by reading a variable in M_{0i0}; the lead processor can obtain a message from a computational processor by a similar read of some M_{i11}. In effect, there are no shared variables.

(d) If a computational processor P_i ($i > 0$) wants to use M_{i10} for "scratch" work, it can close and electronic shutter, temporarily shielding that chip from the constant updating by M_{0i0}.

The motivation for such a configuration can be seen by the following two examples of the use of this system, the first performing at sorting operation, and the second one a Gaussian elimination procedure. *Note that maximal parallelism is obtained in both examples; there is no contention for memory, for shared variables, or for interconnect network resources, and all data transfers (e.g. transfer of subarrays in Example 1) are fully parallel (e.g. a whole subarray can be transferred in one memory cycle).*

Example 1: Sorting

The lead processor will break up the array into subarrays (e.g. as in Quicksort or Mergesort), send them out to the computational processors for sorting, then combine them to produce the sorted version of the original array.

The lead processor P_0 executes the following code

```
form p subarrays in chips of the labeling form
$M_{0i0}$
set a Ready variable in these chips
for i:=1 to p do
    begin
    watch $M_{i11}$ for $P_i$'s Done variable to be set
    read $M_{i11}$ to get the sorted subarray
    end
combine the sorted subarrays, yielding
    the sorted original array
```

P_i executes the following code:

```
watch $M_{0i0}$ for Ready variable to be set
get subarray from $M_{0i0}$
sort subarray
set Done variable in $M_{i11}$
```

Example 2: Gaussian Elimination

Each computational processor is assigned a group of contiguous columns in the matrix. Below we give partial code, showing an operation on a particular diagonal element, say the d-th.

The lead processor P_0 executes:

```
put the value of the divisor (reciprocal of diag elt)
    in a variable in the chips $M_{0i0}$
set a Ready variable in these chips
for i:=1 to p do
    watch $M_{i11}$ for $P_i$'s Done variable to be set
retrieve final matrix from the $M_{i11}$
```

P_i executes:

```
watch $M_{0i0}$ for Ready variable to be set
get divisor from $M_{0i0}$
for all columns assigned to $P_i$ do
    begin
    divide by divisor, yielding w
    for all rows except d do
        subtract w from this (row,col) element
    end
set Done variable in $M_{i11}$
```

Other Architectures:

The above two architectures are just two examples: many other configurations are possible. For example, purely LC systems can be formed, e.g as ring networks. But again, instead of the serial interprocessor communication available in ordinary ring networks, the optical channels introduced here would provide exceedingly highly parallel communication.

Another possibility is to set up memory hierarchies. In this setting, motivated by a desire to conserve on optical apparatus, only some memory access would be optical, with the optically accessed memories serving in the role of cache front ends for much larger memories.

4. Some Implementation Details

Implementation of *OPTIMUL* requires materials and components for the illumination, electro-optic conversion of data, and the subsequent conversion of the data back to an electrical signal. The illumination for the system is provided by a laser at a suitable wavelength and with suitable optical power, as determined by the other components materials in the system. Appropriate optics would be used to focus the beam on the processor imaging arrays. The simplest implementation for *OPTIMUL* involves coating the communication arrays with either advanced, ultrafast, ferroelectric liquid crystals Johnson *et al*, 1987 , or organic thin solid films. Liquid crystal coatings have been applied to integrated circuits in order to create optical diagnostics in order to overcome external *pin limitations in testing* Burns, 1979. *Birefringence* induced by the circuit voltages creates an intensity display in the liquid crystal coating in much the same manner as in digital watches. The nanosecond response of this material translates into data rates of $10^6 b$, 10^{-8} s, approximately 10^{14} b s, for each processor.

Nonlinear organic thin solid film materials are currently of great interest for use in integrated optics Mourou and Meyer, 1984), and studies of the fundamental characteristics of such materials Garito and Singer, 1982) indicate that they posses nonlinear optical figures of merit which are many orders of magnitude greater than inorganic materials such as $LiNbO_3$ and $LiTaO_3$, as well as offering response times approaching femtoseconds. The most likely mechanism to be employed in these films is the linear electro-optic effect, whereby a change in polarization proportional to the chip voltage would be induced. A polarizer inserted above the chip would produce an intensity map of the array gates from the electrically induced birefringence in the film. We are working to deposit such films using spin-coating, as well as by the Langmuir Blodgett technique. The Langmuir/Blodgett method is based on the sequential extraction of molecular monolayers from a liquid surface onto a substrate Kowel *et al*, 1987

Another approach would be to use a material with an *electrochromic property, which would produce an intensity* map of the electrodes directly through electrically induced absorption of light. The Stark effect has been used to characterize L/B films Blinov *et al*, 1984 , and would be an alternative technique which would not require polarizers above the film. Even though such an interaction is likely to be slower than the electro-optic effect, it may be a *feasible implementation since* so large an amount of data is transferred simultaneously.

The deposition of these films should provide excellent topographic coverage, be physically and chemically robust, and be of very uniform thickness and optical quality.

A large number of processors can be accommodated by introducing "fly's-eye" optics capable of imaging the shared memory contents onto a large number of processors, as depicted in Figure 4. CCD arrays are used as receiver/transmitters in that figure, but as mentioned before, DRAM or other technology is possible.

This configuration also allows for broadcast of a system clock from the shared memory, so that all processors can run in a synchronous mode if desired, although they may generate multiple phases or frequencies from the master clock for internal use. The clock could consist of the controller for the Q-switched lasers which serially illuminate the various arrays for their writing opportunities.

5. Acknowledgement

This work was supported in part by the Rome Air Development Center (RADC).

References

D. Agrawal, *Advanced Computer Architecture*, IEEE Computer Society, 1986.

D. Bailey, "Vector Computer Memory Bank Contention *IEEE Transactions on Computers*, 1987, C-36, 3 293-298

T. Bell, "Optical Computing: A Field in Flux," *IEEE Spectrum*, August 1986, 23, 8, 34-37

L. Blinov *et al*, "Polar Langmuir-Blodgett Films, *Thin Solid Films*, 1984, 120, 161-170

J. Bruner, private communication, 1986

D.J. Burns, "Microcircuit Analysis Techniques Using Field-Effect Liquid Crystals," *IEEE Transactions on Electronic Devices*, 1979, ED-26(1), 90-95.

T. Cheung and J. Smith. "A Simulation Study of the Cray X-MP Memory System." *IEEE Transactions on Computers*, 1986, C-36, 7, 613-622.

A. Garito and K. Singer. Organic Crystals and Polymers—A New Class of Nonlinear Optical Materials," *Laser Focus*, 1982, 80, 59-65.

L. Hutcheson, P. Haugen and A. Husain. "Optical Interconnects Replace Hardware," *IEEE Spectrum*, 1987, 24, 3, 30-35.

K. Hwang and F. Briggs. *Computer Architecture and Parallel Processing*, McGraw-Hill, 1984.

K. Johnson, M. Handschy and L. Pagano-Stauffer. "Optical Computing and Image Processing with Ferroelectric Liquid Crystals," *Optical Engineering*, 1987, 26, 385-391.

P. Kogge, *The Architecture of Pipelined Computers*, McGraw-Hill, 1981.

S. Kowel *et al*, "On-Line Diagnostics for Langmuir/Blodgett Film Growth," *Thin Solid Films*, 134, 209-216, 1985.

S. Kowel *et al*, "Future Applications of Ordered Polymeric Thin Films," *Thin Films*, 1987, 377-403.

G. Mourou and K. Meyer, "Subpicosecond Electro-optic Sampling Using Coplanar Strip Transmission Lines," *Applied Physics Letters*, 1984, 45, 492-494.

W. Oed and O. Lange, "Modelling, Measurement and Simulation of Memory Interference in the Cray X-MP," *Parallel Computing, 1986, 343-358*.

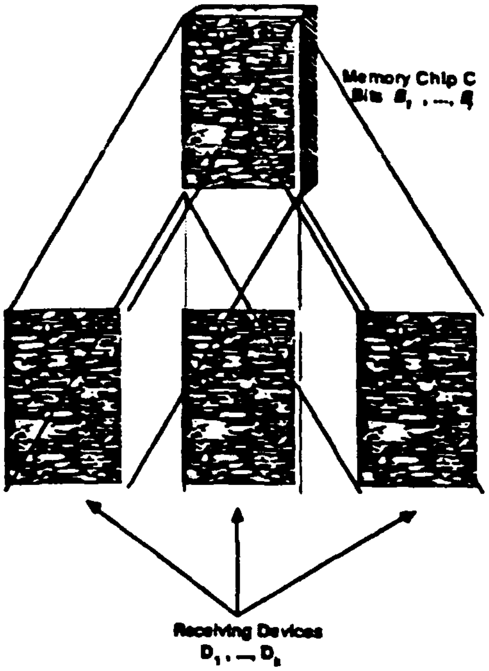
J. Peterson and A. Silberschatz. *Operating System Concepts* (second edition), Addison-Wesley, 1985.

G. Pfister and V. Norton, "Hot spot contention and combining in multistage interconnection networks," *Proceedings of the 1985 International Conference on Parallel Processing*.

D. Siewiorek, C. Bell and A. Newell. *Computer Structures: Principles and Examples*, McGraw-Hill, 1982.

A. Wilson, "Hierarchical cache/bus architecture for shared memory multiprocessors," *Proceedings of the 14th Annual International Symposium on Computer Architecture*, 1987, 244-252.

Optical Transfer of Information from Main Memory to Local Memories



Illumination of memory chip C allows simultaneous transfer of bits B0-Bk to all receiving devices. For a memory size of 1K and a transfer time of 100 ns the effective data rate is greater than 10GBits/sec.

FIGURE 1

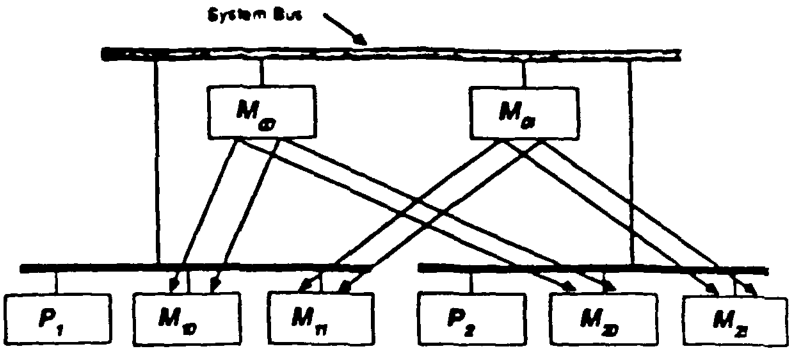


FIGURE 2

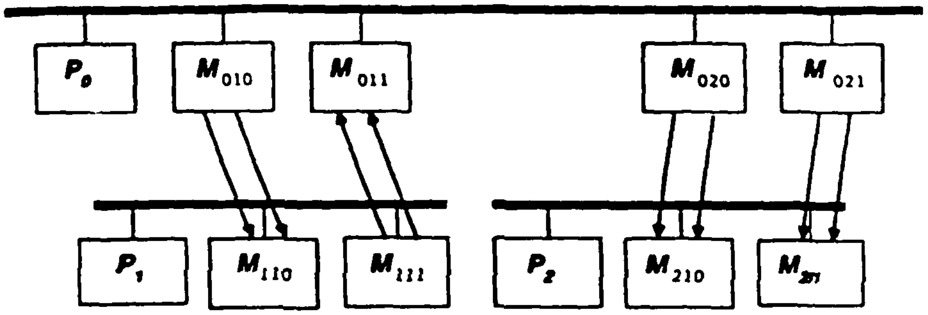


FIGURE 3

OPTIMUL

FLY'S - EYE LENS DISTRIBUTION

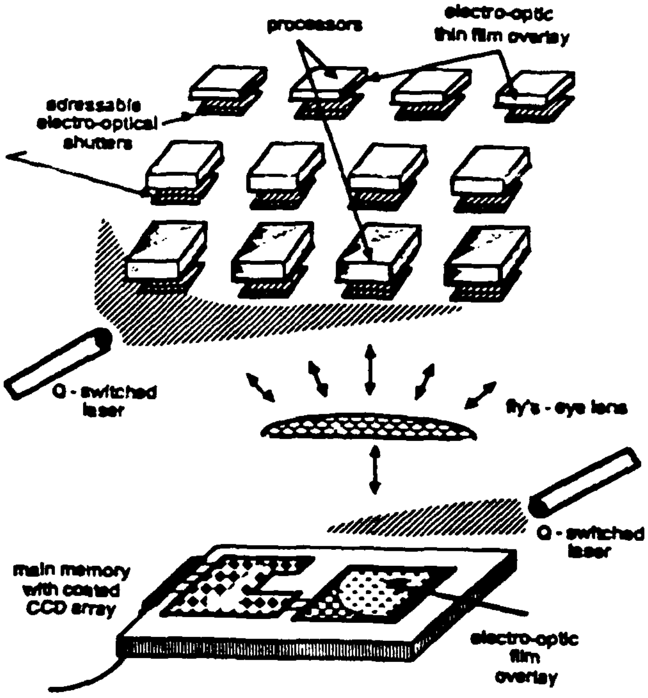


FIGURE 4

To appear in *Proceedings of the IEEE Phoeniz International Conference on Computers and Communications*, March 1989.

Performance Analysis of the OPTIMUL Multiprocessor Interconnect

N. Matloff, T. Schubert, S. Kowel, C. Eldering, M. Loving

Department of Electrical Engineerıng & Computer Science

University of California at Davis

1. Introduction

Multiprocessor (MP) systems, consisting of p interconnected but independent processors, have the *potential* for a speedup factor of p in computational power. However, a long-standing problem has been that this potential has not been realizable, due to the overhead of processor-memory and/or processor-processor communication. This has been the case for both types of MP systems which are usually considered:

Tightly-Coupled (TC) Systems:

The very significant overhead in TC systems takes the form of contention for the shared central memory M_{cent}, and for the processor-memory interconnect. The latter problem is exacerbated by the fact that the expense of full crossbar switches results in the use of other networks for which there is even more processor contention for the interconnect, e.g. Ω*-nets* [Hwang and Briggs, 1984]. For small values of p, use of caches can be effective, but the efficiency decreases with p [Wilson, 1987]. Furthermore, it has recently been discovered that access to interprocess synchronization variables in shared memory worsens this problem tremendously [Pfister and Norton, 1986].

Loosely-Coupled (LC) Systems:

In LC systems, there is no shared memory, but there is still communications overhead of another kind. The processors communicate with each other through a network. Bandwidth limitations on this interconnection network present very substantial overhead. For example, intercluster accesses in the Cm* machine were a factor of 8.7 times slower than accesses to local memory [Hwang and Briggs, 1984].

In both the TC and LC settings, another significant problem is the severe restrictions resulting from chip pin limitations. Even channels of very high bandwidth, such as those constructed from optical fibers, would not solve the problem arising from the fact that there are only a few data pins but thousands or even millions of bits in a memory chip.

The entire history of the development of MP technology has been dominated by the search for solutions to these problems [Siewiorek *et al*, 1982; Hwang . nd Briggs, 1984; Agrawal, 1986]. Essentially, no completely satisfactory solutions have been found. For example, after Cray Research, Inc. released the Cray X-MP, an MP version of the Cray-1 supercomputer recently, a number of investigations [Bailey, 1987; Cheung and Smith, 1986; Oed and Lange, 1986] quickly showed the system to suffer from slowdowns due to both contention for shared memory and contention for the network which connects the processors to that memory, just as with all the earlier MP systems.

Perhaps an even more dramatic example is the S-1, a TC MP system developed at Lawrence Livermore National Laboratories [Hwang and Briggs, 1984]. Throughout the period of development of this system, it was hailed as one of the most advanced MP projects in existence. However, recently the project was discontinued, in spite of all the favorable publicity, and the very extensive funds expended [Bruner, 1987]. One of the primary reasons given for the discontinuation was that the project engineers had found that the contention for shared memory in the system would be much greater than they had anticipated. They are now beginning work on a completely new design.

Such problems have been considered extremely difficult to solve, with some authors even going so far as to say that we possibly should resign ourselves to the problems not being solved, concentrating on software methods instead [Ledbetter, 1988].

However, in [Matloff, Kowel and Eldering, 1988] a radically new interconnect method was presented which will solve these problems, and have other advantages as well: The new interconnect will be usable for both fine-grained and coarse-grained types of applications; it will solve the long-standing problems of contention for memory and for the processor/memory switch in TC systems; in LC systems, it will enable a truly dramatic improvement in interprocessor communications bandwidth, and again totally eliminate contention for the interconnect switch; our approach should also be superior to other optical processor/memory interconnects which have been proposed, e.g. optical crossbars [Bell, 1986; Hutcheson *et al*, 1987]; the pin-limitation problem is also eliminated, which is a problem even in those architectures which have been proposed based on an optical fiber interconnect.

Our name for this new interconnect is OPTIMUL, an acronym for Optical Multiprocessor Interconnect. The central feature is an optical processor-memory channel, which will allow simultaneous access of a memory chip, where the word "simultaneous" is meant *both* with respect to all bits in the chip, *and* with respect to all processors. In other words, all processors can simultaneously read the entire contents of a chip, with no interference at all. Write access is of course restricted to a single processor at a time, but it still is simultaneous across bits in the chip, i.e. an entire chip can be written in one access. This optical channel is described in Section 2, and then MP system architectures utilizing it will be proposed in Section 3. Sections 4 and 5 will present some performance analyses of these architectures.

2. A New Optical Memory Access Channel

Consider devices D_1, ..., D_k which wish to read a memory chip C, in which are stored bits B_1, ..., B_r. We will report here a technique in which the devices can read from C optically, bypassing the need for using the chip's pins, and which will allow this access to be simultaneous, with respect to both devices D_i and bits B_j (Figure 1).

To achieve this, C will be illuminated and mechanisms used to cause the reflected light beam to be intensity modulated by the electric fields at each position in C. Thus the reflected beam will contain a complete bit map of the contents of C. The beam will be demodulated by optical apparatus for focusing onto the receivers D_j.

Some preliminary implementation details were given in [Matloff, *et al*, 1988]. An updated is given in the following:

To achieve the desired modulation effect, we are pursuing two strategies, one based on advanced ferroelectric liquid crystals, and the other using thin solid film structures containing highly nonlinear dyes. Either material would be used to coat over the surface of the chip C above (or a group of such chips).

The fields on the surface of typical IC's are of magnitude on the order of volts/μm, larger than the fields supplied by the electrodes in a typical liquid crystal display. This fact led to the demonstration of an electro-optical method for testing integrated circuits [Burns, 1979]. Problems such as long switching times (~ 10 ms) have recently been resolved, with switching times on the order of 100 ns, and even faster operation appears to be possible [Johnson, *et al*, 1987].

We also have been examining the feasibility of using thin solid organic films as the coating material to be used to effect the light modulation. Such materials appear promising, and would offer a tradeoff of higher speed for lower image contrast [Kowel, *et al*, 1987] [Kowel, 1985]. We are investigating synthesis and deposition techniques, and are collecting electro-optical measurements to evaluate the potential of these films.

Demodulation of the beam back to storage as electric fields at the receivers is to be accomplished by the use of photosensitive technology. For example, one possibility is to use ordinary DRAM memories, which have a natural sensitivity to light. This means also that parts of C must be masked from the light, so that illumination of C does not change the contents of bits in C; e.g. only the output portion of a gate can be exposed.

However, in commercially produced chips, this photosensitivity of DRAM's may not be uniform enough for reliable use as demodulators, since the sensitivity is a byproduct, not a primary specification. Thus, we are taking other approaches instead, based on photodiodes. We have designed and simulated such a receiving device [Loving and Eldering, 1988]. In fact, other such memories have been proposed [Kosnocky, 1971] [Ullman *et al*, 1988].

In this way, the values stored at all the bits B_j in C can be transmitted optically to the devices D_i, *simultaneously over all subscripts i and j*. Clearly, the simultaneity over i will have highly significant implications for the memory and interconnect contention problems which have plagued TC systems, while the simultaneity over j will have an equally profound impact on the network bandwidth limitation problem in LC systems. Note that the classical bottleneck arising from limitations on the pins-to-stored-bits ratio is completely bypassed in the approach described here. Writes to C can be accomplished by reversing the process.

3. System Architectures

The optical interconnect presented here can be used in a variety of configurations. Two of these were described in [Matloff, Kowel and Eldering, 1988], which will be summarized here:

Architecture I:

This configuration features optical memory reads, but used electronic writes, the latter being motivated by a desire for simplicity in the first prototype to be constructed, and by the fact that the electronic bus, with a standard Test-and-Set cycle or similar mechanism, avoids the interprocess synchronization problem which must be solved in a purely optical system.

Architecture II:

This configuration features both optical reads and writes. It is intensive in memory quantity needed, with essentially separate memory modules being used for reads and writes. Interprocess synchronization is handled by message-passing techniques [Peterson and Silberschatz, 1985], the implementation of which were given in examples in [Matloff, Kowel and Eldering, 1988].

Other Architectures:

The above two architectures are just two examples; many other configurations are possible. For example, purely LC systems can be formed, e.g. as tree or ring networks (see Section 5). But again, instead of the serial interprocessor communication available in ordinary ring networks, the optical channels introduced here would provide exceedingly highly parallel communication. This is currently being investigated [Matloff and Schubert, 1988].

Another possibility is to set up memory hierarchies. In this setting, motivated by a desire to conserve on optical apparatus, only some memory access would be optical, with the optically accessed memories serving in the role of cache front ends for much larger memories.

4. Performance Analysis: Simulation of a Continuum of Systems with Varying Degrees of Coupling

Numerous mathematical analyses of multiple access of memory systems have been presented (a nice collection of references appears in the introduction to Chapter 6 of [Agrawal, 1986]). However, for the present purpose, a simulation analysis was preferred, in the interests of (a) simplicity, and (b) modeling OPTIMUL's ability of a processor to do a parallel access of a large data structure.

Specifically, we set up the following model, which can be considered as an abstraction which is representative of a number of architectures which could be developed using the optical interconnect introduced in [Matloff, Kowel and Eldering, 1988]. We will refer to the abstracted system here by the same name, OPTIMUL.

In this system we have p processors viewing a central shared memory of m modules. Consider the operation of one processor P. P will alternate between periods of memory access and nonaccess. We assume the nonaccess time (measured in units of memory cycles) has a geometric distribution with mean μ_{nonacc}.

The model then assumes that at the start of an access period, P will send to a memory controller a request for R_{wrds} consecutive words in the memory space, e.g. a request to read an entire array or subarray. R_{wrds} is assumed to have a geometric distribution with mean μ_{wrds}.

We are comparing OPTIMUL to a conventional MP system. There is extremely wide variation in "conventional" MP systems; the model cannot incorporate all of them. Instead the model has been designed so that variation of its parameters will allow modeling of a range of situations suitable for comparison to OPTIMUL; this will be seen below.

In the conventional system, it is assumed that the memory controller will satisfy the R_{wrds} requests made by P in whatever order they become satisfiable, similar to the "C" organization [Hwang and Briggs, 1984] [Kogge, 1981], with consecutive words stored in consecutive modules (mod m), i.e. using low-order interleaving. If one of the words requested by P encounters contention with a request from another processor, one of the processors *must wait. If a requested module is free*, it takes one unit of time to satisfy a request for one word of memory.

On the other hand, in modeling OPTIMUL, we are assuming that *any* request takes only one unit of time to service, for any value of R_{wrds}, i.e. OPTIMUL will access *all* R_{wrds} in one time unit, due to OPTIMUL's ability to transfer the entire contents of a memory chip in parallel. [For this reason, this way would be most fully exploited if the M_{loc}'s (as in Architecture I) were contained within the processors, we are making such an assumption here. On the other hand, in some ways our model is too conservative, i.e. it actually underestimates OPTIMUL's potential; this will be explained below.

The simulation actually measures the performance of our model, conventional MP system, rather than OPTIMUL itself. The mean delay per memory access, D_C, is found for the conventional system. Under the model described here, the corresponding mean delay for OPTIMUL is exactly 1.0. Thus D_C may be used as a figure of merit for OPTIMUL, i.e. a measure of the speedup in memory access obtained.

We have noted that one of OPTIMUL's important advantages is that it can operate in both TC and LC modes. This is the motivation behind our model for the memory access of a conventional MP:

TC Model:

We model the "typical" TC system as having a fairly small value of μ_{nonacc}, 2.0. This reflects the fact that TC systems are appropriate for applications in which the processors must communicate with each other fairly often, and that they do so by accessing M_{cent}. However, such accesses are usually for only one word, or a small number of words. To reflect this, we set μ_{words} to be fairly small in our simulator. To model TC systems which exist today, in which the number of processors is limited, we will set the number of processors p to be small in our TC simulations, specifically 16.

LC Model:

Here we set p to be a larger number (64) in our simulator, reflecting the situation in many current LC systems (of course, many such systems are even larger than this). Also, since LC systems are set up for applications in which the processors communicate with each other less frequently, we have set μ_{nonacc} to be fairly large (100.0). However, when LC systems do communicate with each other, it tends to be with relatively large amounts of data; thus we have set μ_{words} to be fairly large in our simulation, with a value of 100.0.

Both the TC and LC models include m = 16 memcry modules in M_{cent}.

Note that these models will severely *underestimate* *OPTIMUL*'s potential, in a number of ways. For example, the TC model tacitly assumes that the processor/memory interconnect switch for the conventional MP machine is in the form of a crossbar, which is not typical in MP systems, and is actually infeasible for the larger ones. Thus the model for the conventional MP machine does not incorporate any queueing delay due to the interconnect switch; as mentioned above, such delay can be quite large, and thus this results in underestimating *OPTIMUL*'s potential. Of course this built-in bias against OPTIMUL will be even worse in our LC model, since the interconnect queueing delay is much worse in that case; we are not allowing for network traffic delay at all in this simple analysis.

A large number of simulation runs were conducted, but instead of reporting all of them, we will concentrate on three representative examples:

Example A:

This is a TC model, with $\mu_{words} = 1.0$. This setting can be expected to give only a modest advantage to OPTIMUL over conventional machines, due to the above-mentioned lack of interconnect queueing delay in our model. However, we still found that the figure of merit D_C was 1.34, i.e. even this setting's bias against OPTIMUL, OPTIMUL has a 34% advantage.

Example B:

This too is a TC model, but with $\mu_{words} = 10.0$, representing a situation in which the P_i are vector processors. This models a setting in which most memory accesses of a processor are for scalars, but occasionally a vector access is made. Here we found that $D_C = 12.44$, a 12-fold advantage for OPTIMUL.

Example C:

> This is the LC model described above. Here OPTIMUL has a very dramatic advantage over a conventional system, with $D_C = 277.35$ (and, as mentioned above, this number is probably an underestimate of the true value).

In addition, one of OPTIMUL's most significant advantages is invisible in the simulation study, namely the feasibility of using a much larger number p of processors in a TC system. The limitations of crossbars (or their more sophisticated variations) on p imply that it would be infeasible to use TC systems in applications having a very high degree of inherent parallelism. The optical interconnect nature of OPTIMUL should make it much more feasible to build large TC systems, so that more highly parallel applications may be handled.

5. Performance Analysis: Case Study of a Sorting Application

In Section 4, we presented an analysis based on abstraction of memory access patterns in multiprocessor systems. This analysis showed the potential of OPTIMUL to be quite dramatic for some settings of the simulation parameters. However, additional understanding is gained by investigating the performance of OPTIMUL on a specific application, which is done in this section. The analysis here is basically a trace-driven simulation of the performance of our proposed system on sorting problems.

The analysis assumes that OPTIMUL's processors are of speed comparable to that of a *VAX 8600. Single-processor computation times used* below were obtained by using the Unix 'time' command to get processor run times for actual C code for the sort algorithm specified *below*.

The processors are assumed to be set up as an LC system in a ring topology. The OPTIMUL version of this system is assumed to have optical neighbor-to-neighbor links which use the technology described above, which the capability of transferring millions of bits in hundreds of nanoseconds; interprocessor communication time is essentially negligible in this system. The non-OPTIMUL version of the system has "conventional" neighbor-to-neighbor links having transmission rates of 50 megabits/second. Links of this speed or better are beginning to appear, e.g. the "semi-LC" VAX Cluster sytems [Kronenberg, Levy and Strecker, 1986], this rate is much faster than is typical among most LC systems to date, e.g. the Hypercube.

The sort algorithm used was Quickmerge [Quinn, 1987], which consists of three phases. During Phase I each processor sorts a subset of the array using Quicksort. These subsets must then be merged to complete the sort. Before the merge phase, Phase II, a search phase, Phase III, is added so that the merge task car be divided among all the processors. Processors search for dividers to partition each of the sorted subsets such that there is no value in $\text{partition}_{i,j}$ of any subset j which is greater than any value in $\text{partition}_{i+1,k}$ of any subset k. During the merge phase, each processor joins together a set of partitions which share common dividers. Because of the divisions performed in step two, merged part.tion$_i$ precedes merged partition$_{i+1}$.

On an LC system, the communication between phases is substantial:

(a) Before the initial (sort) phase, each processor must receive a subset of the array to sort. These subsets are sent by the lead processor, relayed from processor to processor along the ring until reaching the desired destination processor.

(b) Before the second (search) phase begins, each processor must receive the sort phase results from all other processors.

(c) Before the final (merge) phase, the partition dividers must be passed to each of the processors (note that the data is already present in each processor's private memory).

(d) Finally, the merged partitions must be returned to the lead processor for concatenation.

The entire array must be broadcast three times. As the total communications cost is dominated by this data movement, we won't consider the transfer time of the partition dividers.

Within an OPTIMUL ring configuration, memory would appear to be shared since information could be transfered continuously around the ring. As OPTIMUL allows a complete memory-memory transfer in one memory cycle, data can be transferred (broadcast) to all processors in p memory cycles where p is the number of processors on the ring. Preliminary study suggests that we will be able to transfer the contents of one memory chip to another in less than 500ns, and that this time can be reduced to less than 100ns. Even given the slower speed, data could be broadcast to all members of a 64 member ring in about 32μs (63*500ns = 31500ns << 1 ms). This is a substantial savings over the alternatives discussed above, even ignoring the propagation delay around the ring.

Below are tables indicating approximate times for the Quickmerge algorithm were the algorithm executed on OPTIMUL and non-OPTIMUL ring as described above. The improvements look modest in comparison with that of Example [illegible] in the last section, but still are quite impressive, with speeds double and triple those of the conventional LC system. The largest improvement reported occurs for a 128 processor system sorting a 256k integer array. Here the OPTIMUL system would perform approximately three and a half times faster than a non-OPTIMUL system having the same number of processors. For larger problems and more processors, larger speedup factors might be observed. [On the other hand, it appears that additional tuning of the algorithm could be done for the non-OPTIMUL setting, and the gap in performance narrowed somewhat.]

More detailed analyses, including the implementation details for such a ring configuration, are currently in progress [Matloff and Schubert, 1988].

The gains reported here are significant, but modest in comparison to the most extreme gains presented in Section 4. In that light, it must once again be pointed out that speedup factors are highly application-dependent. In particular, in the sorting application analysed here, there is a fundamental obstacle to speedup, in terms of the relative size of computation and communication times:

Consider sorting n items on p processors, by partitioning into blocks of approximate size n/p each. The computation time is approximately C n/p log(n/p) for some C, assuming that all subproblems finish at roughly the same time; this is not a bad assumption, since the standard deviation of sort times is small compared to the mean [Gonnet, 1984]. [For simplicity, we are ignoring the merge phase in the analysis below.] The communications time is roughly D (n/p) p (n/p amount of data being passed through p nodes) for some D.

Fix p and vary n. If the ratio n/p is too small, then very little data is being passed from node to node, not enough to fully exploit the highly parallel data transmission capability in OPTIMUL. On the other hand, as n grows, the computation time tends to dominate the communication time. In this setting, OPTIMUL's communications advantages will be quite substantial over non-OPTIMUL systems, but the advantages will not be important, since communications times will be a minor proportion of the total times anyway.

In other words, applications such as sorting, having computation times which are more than O(n), are poor candidates for studies whose aim is to investigate interprocessor communications costs. In such applications, inefficient interprocessor communication might not be penalized much. Searching applications, with O(n) or O(log n) computational times, should much more fully exploit OPTIMUL's hugh communications bandwidth capabilities, and are currently under investigation.

6. Acknowledgement

This work was supported in part by Rome Air Development Center (RADC).

Tables

Time (ms) to sort 64k integer array
communications time per transfer: 40.00ms (50.0Mbit/sec)

# of	SORT		SEARCH		MERGE		TOTAL		
processors	min	max	min	max	min	max	OPTIMUL	nonOPTIMUL	speedup
1	-	2672	-	-	-	-	2672		
2	1232	1264	0	0	864	880	2144	2264	1:1.06
4	560	608	0	1	560	672	1281	1401	1:1.09
8	240	288	0	1	304	352	641	661	1:1.22
16	96	144	1	2	128	192	338	458	1:1.36
32	48	80	2	3	48	80	163	283	1::174
64	16	48	4	6	16	48	102	222	1:2.18
128	0	32	9	13	0	32	77	197	1:2.56

Time (ms) to sort 128k integer array
communications time per transfer: 80.00ms (50.0Mbit/sec)

# of	SORT		SEARCH		MERGE		TOTAL		
processors	min	max	min	max	min	max	OPTIMUL	nonOPTIMUL	speedup
1	-	5840	-	-	-	-	5840		
2	2624	2736	0	0	1760	1856	4592	4592	1:1.05
4	1232	1264	0	1	1184	1200	2465	2705	1:1.10
8	560	672	0	2	688	704	1378	1618	1:1.17
16	224	304	1	2	288	400	708	948	1:1.34
32	96	144	2	4	128	160	308	548	1:1.78
64	32	80	5	7	48	80	167	407	1:2.44
128	0	48	9	15	16	80	143	383	1:2.68

Time (ms) to sort 256k integer array
communications time per transfer: 160.00ms (50.0Mbit/sec)

# of	SORT		SEARCH		MERGE		TOTAL		
processors	min	max	min	max	min	max	OPTIMUL	nonOPTIMUL	speedup
1	-	11952	-	-	-	-	11952		
2	5536	5664	0	0	3456	3472	9136	9616	1:1.05
4	2624	2640	0	1	2368	2464	5105	5595	1:1.09
8	1200	1280	0	2	1376	1568	2850	3330	1:1.17
16	544	608	1	2	656	704	1314	1794	1:1.37
32	224	336	3	4	288	368	708	1188	1:1.68
64	96	160	5	8	112	176	344	824	1:2.40
128	32	80	10	17	48	96	193	673	1:3.49

References

D. Agrawal, *Advanced Computer Architecture*, IEEE Computer Society, 1986.

D. Bailey, "Vector Computer Memory Bank Contention," *IEEE Transactions on Computers*, 1987, C-36, 3, 293-298.

T. Bell, "Optical Computing: A Field in Flux," *IEEE Spectrum*, August 1986, 23, 8, 34-37.

J. Bruner, private communication, 1986.

D. Burns, "Microcircuit Analysis Techniques Using Field-Effect Liquid Crystals," *IEEE Transactions on Electronic Devices*, 1979, ED=26(1), 90-95.

T. Cheung and J. Smith, "A Simulation Study of the Cray X-MP Memory System," *IEEE Transactions on Computers*, 1986, C-36, 7, 613-622.

G. Gonnet, *Handbook of Algorithms and Data Structures*, Addison-Wesley, 1984.

L. Hutcheson, P. Haugen and A. Husain, "Optical Interconnects Replace Hardware," *IEEE Spectrum*, 1987, 24, 3, 30-35.

K. Hwang and F. Briggs, *Computer Architecture and Parallel Processing*, McGraw-Hill, 1984.

K. Johnson, *et al*, "Optical Computing and Image Processing with Ferroelectric Liquid Crystals," *Optical Engineering*, 1987, 26(5), 385-391.

P. Kogge, *The Architecture of Pipelined Computers*, McGraw-Hill, 1981.

S. Kowel *et al*, "On-Line Diagnostics for Langmuir/Blodgett Film Growth," *Thin Solid Films*, 134, 209-216, 1985.

S. Kowel *et al*, "Future Applications of Ordered Polymeric Thin Films," *Thin Films*, 1987, 377-403.

W. Kosnocky, "Electrically and Optically Accessible Memory," U.S. Patent #3,631,411, December, 1971.

C. Leadbetter, invited lecture, 1988 ACM International Conference on Supercomputing, Saint-Malo, France.

M. Loving and C. Eldering, "Design of a Receiver/Memory Chip for an Optical Multiprocessor Interconnect," Technical Report, Division of Computer Science, University of California at Davis, 1988.

N. Matloff, S. Kowel and C. Eldering, "OPTIMUL: An Optical Interconnect for Multiprocessor Systems," *Proceedings of the 1988 ACM International Conference on Supercomputing*, 14-24, Saint-Malo, France.

N. Matloff and T. Schubert, "Topologies for Optically Interconnected Multiprocessor Systems," work in progress.

W. Oed and O. Lange, "Modelling, Measurement and Simulation of Memory Interference in the Cray X-MP," *Parallel Computing*, 1986, 343-358.

J. Peterson and A. Silberschats, *Operating System Concepts* (second edition), Addison-Wesley, 1985.

G. Pfister and V. Norton, "Hot spot contention and combining in multistage interconnection networks," *Proceedings of the 1985 International Conference on Parallel Processing.*

D. Siewiorek, C. Bell and A. Newell, *Computer Structures: Principles and Examples*, McGraw-Hill, 1982.

Ullman, *et al*, "Method and Apparatus for Loading Information into an Integrated Circuit Semiconductor Device," International Patent Application #PCT/GB85/00404, International Publication #W086/01931, March 1986.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control, Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of competence is provided to ESD Program Offices (POs) and other ESD elements to perform effective acquisition of C^3I systems. The areas of technical competence include communications, command and control, battle management information processing, surveillance sensors, intelligence data collection and handling, solid state sciences, electromagnetics, and propagation, and electronic reliability/maintainability and compatibility.